Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

Protein folding by distributed computing and the denatured state ensemble

Neelan J. Marianayagam, Nicolas L. Fawzi, and Teresa Head-Gordon

PNAS 2005;102;16684-16689; originally published online Nov 2, 2005;
doi:10.1073/pnas.0506388102

Protein folding by distributed computing and the denatured state ensemble

Neelan J. Marianayagam*[†], Nicolas L. Fawzi[†][‡], and Teresa Head-Gordon*[‡][§][¶]

*Department of Bioengineering and [‡]UCSF/UCB Joint Graduate Group in Bioengineering, University of California, Berkeley, CA 94720;
and [§]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Edited by Alan R. Fersht, University of Cambridge, Cambridge, U.K., and approved September 27, 2005 (received for review July 26, 2005)

The distributed computing (DC) paradigm in conjunction with the folding@home (FH) client server has been used to study the folding kinetics of small peptides and proteins, giving excellent agreement with experimentally measured folding rates, although pathways sampled in these simulations are not always consistent with the folding mechanism. In this study, we use a coarse-grain model of protein L, whose two-state kinetics have been characterized in detail by using long-time equilibrium simulations, to rigorously test a FH protocol using ≈10,000 short-time, uncoupled folding simulations starting from an extended state of the protein. We show that the FH results give non-Poisson distributions and early folding events that are unphysical, whereas longer folding events experience a correct barrier to folding but are not representative of the equilibrium folding ensemble. Using short-time, uncoupled folding simulations started from an equilibrated denatured state ensemble (DSE), we also do not get agreement with the equilibrium two-state kinetics because of overrepresented folding events arising from higher energy subpopulations in the DSE. The DC approach using uncoupled short trajectories can make contact with traditionally measured experimental rates and folding mechanism when starting from an equilibrated DSE, when the simulation time is long enough to sample the lowest energy states of the unfolded basin and the simulated free-energy surface is correct. However, the DC paradigm, together with faster time-resolved and single-molecule experiments, can also reveal the breakdown in the two-state approximation due to observation of folding events from higher energy subpopulations in the DSE.

folding mechanism | folding@home | two-state kinetics | Poisson process | coarse-grain model

As simulation and experimental folding timescales begin to coincide, a much more direct comparison between experiments and simulations is finally possible. This overlap has been a particular strength of the distributed computing (DC) approach known as folding@home (FH), which compares simulation results against experiments performed on the same protein to understand how and how fast simple proteins fold (1–5). The DC approach, based on the two-state approximation (6), assumes that for proteins that fold through a single, rate-limiting barrier, the folding population follows a Poisson distribution so that the time between folding events follows

$$1 - \exp(-kt), \qquad [1]$$

which can be expanded as a MacLaurin series

$$1 - \left[1 - kt + \frac{1}{2}(kt)^2 + \ldots\right], \qquad [2]$$

so that the fastest folding trajectories dictate the slope and therefore the kinetic rate constant k of the first-order term in Eq. **2**. Although these fastest folding events are relatively rare (≈10 for every 10,000 trajectories that fold within a simulation time t_{sim}), the DC approach is ideally suited for generating tens of thousands of short-timescale and independent trajectories to determine these far fewer folding events. We distinguish the DC approach, defined as short trajectories starting from a properly equilibrated denatured state ensemble (DSE), from a FH protocol that launches these short trajectories from states that are highly extended protein configurations as a model of the DSE. In both cases, the sets of folding trajectories are uncoupled. It should be noted that the developers of FH use other simulation protocols than the particular version that is analyzed here (1, 7).

The FH protocol of short uncoupled trajectories starting from an extended state has been analyzed by a number of experimental and theoretical groups. Fersht (8) has asserted that simulations started from completely extended conformations are unphysical states under folding conditions so that lag times and sampling of biased pathways will result. Although the folding kinetics of the villin headpiece using FH agreed with the experimental rate (5), Eaton and coworkers (9) have shown that amino acids predicted by FH to be involved in the transition state had no effect on the experimental rate when mutated. Caflisch and coworkers (10), using all-atom simulations of a 20-residue β-hairpin peptide in implicit solvent, show that the FH approach correctly predicts the equilibrium folding time for short trajectories (>1 ns); however, for very short trajectories (<1 ns), the predicted folding time is longer than the equilibrium folding time, the kinetics in this region are non-Poisson, and the mechanism does not agree with equilibrium simulations. Recent theoretical work using Markov models to describe protein-folding kinetics also shows that an internal relaxation time from an artificially created extended state, which is comparable to the equilibrium folding time, will not generate meaningful results (11).

This work carefully examines the FH simulation protocol using uncoupled short trajectories separately from the DC approach for generating kinetic rates and mechanisms and therefore the ability to connect to real experimental observations. We use a comprehensively characterized off-lattice protein model that we developed to study the folding kinetics and thermodynamics of protein L, found by our model (12–14) and by experiment (15) to be a classic two-state folder. Our long-timescale equilibrium folding study has determined the folding temperature, DSE, transition state ensemble (TSE), and kinetic rate for the folding of protein L and therefore serves in this work as the experimental benchmark (12–14). With the same model, we follow the FH protocol described above, and a DC approach that launches short uncoupled trajectories from the true DSE, to determine whether kinetic rates and the folding mechanism conform to that found by "experiment."

Methods

We briefly give some details of the model here. The 20-residue amino acid alphabet is mapped onto three-letter code: B, hydro-

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: DC, distributed computing; FH, folding@home; DSE, denatured state ensemble; TSE, transition state ensemble.

[†] N.J.M. and N.L.F. contributed equally to this work.

[¶] To whom correspondence should be addressed at: Department of Bioengineering, University of California, 204 Calvin Laboratory, Berkeley, CA 94720-1762. E-mail: tlhead-gordon@lbl.gov.

www.pnas.org/cgi/doi/10.1073/pnas.0506388102

phobic; L, hydrophilic; and N, neutral. The details of the mapping and the potential energy function can be found in ref. 12. The system is simulated by using constant-temperature Langevin dynamics with a (low) friction coefficient of $0.05\tau^{-1}$, where τ is the time in reduced units. We also performed Langevin dynamics simulations at a higher viscosity parameter of $0.20\tau^{-1}$ corresponding to the high friction limit. Bond lengths are held rigid by using the RATTLE algorithm (16). All simulations carried out in this work use an optimally designed protein L sequence where the energy gap between the native state and the misfolded structures is maximized (12). The folding temperature, kinetic rate, DSE, and TSE have all been characterized by long equilibrium folding simulations [10 million Langevin time steps ($50{,}000\tau$) and ≈1,000 trajectories] reported elsewhere (12–14). To ensure that our system is robustly two-state, we also have carried out in this work even longer simulations at the folding temperature where we can observe multiple folding and unfolding events.

To analyze the DC approach, we use a fraction of the equilibrium timescale. All of the short-timescale computations were run as independent trajectories on both our in-house G5 Macintosh dual processors and G5 Macintosh cluster. We launch trajectories for 500,000 Langevin time steps (2500τ), which is 5% of the time steps used in the equilibrium simulations and which we define as t_{sim}. The initial states for these short trajectories are prepared depending on the protocol examined. For the DC approach, we use starting structures from the previously reported simulations of our equilibrated DSE. We then launch trajectories from this equilibrated DSE at the folding temperature (T_{f}) of 0.42 and collect first passage times for events that fold within t_{sim}. For the FH protocol, we subject the native configuration of protein L to a large number of high-temperature decorrelation steps to ensure that the starting structures are extended or random-walk states and to serve as a FH estimate of the unfolded state. We then drop the temperature to $T_{\mathrm{f}} = 0.42$ and collect first passage times for those trajectories that fold within t_{sim}. We define a third set of simulations in which we use the trajectories that did not fold from the FH protocol as a new estimate of the DSE (DSE estimate) and collect first passage times for those trajectories that fold within t_{sim}.

We define a structure being in the native basin of attraction by the structural similarity parameter χ.

$$\chi = \frac{1}{M}\sum_{i,j\geq i+4}^{N} h(\varepsilon - |r_{ij} - r_{ij}^{\mathrm{native}}|), \qquad \textbf{[3]}$$

where the double sum is over beads on the chain, and r_{ij} and r_{ij}^{native} are the distances between beads i and j in the state of interest and the native state, respectively. h is the Heaviside step function, with $\varepsilon = 0.2$ to account for thermal fluctuations away from the native state structure. M is a constant that satisfies the conditions that $\chi = 1$ when the chain is identical to the native state and $\chi \approx 0$ in the random-coil state. For $\chi > 0.4$, this protein is said to be in the native state. When $\chi < 0.4$ and the radius of gyration, R_{g}, satisfies $2.5 < R_{\mathrm{g}} < 4.5$, the protein resides in the unfolded basin. We also carried out simulations under the FH protocol where the native state is defined more stringently as $\chi > 0.5$, $\chi > 0.6$, and $\chi > 0.7$, and we found that this definition had no effect on our results below (see the supporting information, which is published on the PNAS web site).

The contact maps shown in Figs. 2 and 3 show which areas of the regions of protein L are in contact in the various states of the protein examined during the folding reaction. Two beads are said to be in contact if they are within 2.5 distance units of each other (≈9 Å), which is the center of mass distance between side chains. In all figures, native contacts are always in black. Contacts that form in the TSE, DSE, or examined points along the folding pathway are contoured at different levels in Figs. 3–5. The TSE is contoured at 80% of the population (blue contours), the DSE is contoured at 45% (red contours), and other points along the folding pathway are contoured between 30% and 50% (green contours).

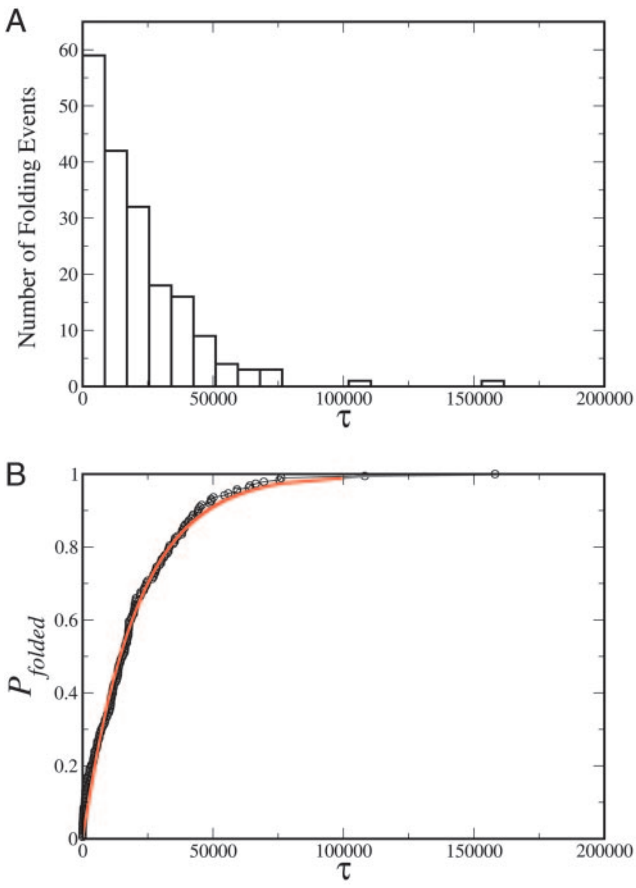

Fig. 1. Distribution of the first passage time (*A*) and plot of percent folded vs. time (*B*) for the long time simulations. The Poisson distribution in *A* and the single exponential fit of *B* show that our model of protein L is a two-state folder.

The results presented in the next section are based on statistics collected from 10,660 uncoupled trajectories launched using the FH protocol starting from highly extended states, from 11,337 uncoupled trajectories launched using the DC approach starting from the estimate of the DSE based on nonfolding FH results, and from 8,046 uncoupled trajectories launched using the DC approach starting from the equilibrated DSE. Of the total FH trajectories, 6,211 trajectories folded (≈58%); for the DC simulations based on a DSE estimate, 3,009 folding events were observed (≈27%); for the DC simulations based on the true DSE, 1,472 folding events were counted (≈18%). We also launched 3,715 FH simulations using the higher viscosity parameter and found that 1,661 trajectories resulted in folding events (≈52%).

Results

In Fig. 1, we show the distribution of first passage times and the resulting cumulative distribution for the long time simulations. These plots show that our system is robustly a two-state folder, with time between folding events following a Poisson distribution and conforming to single exponential kinetics (fit parameters are reported in Table 1).

The analogous plots for the FH simulations are shown in Fig. 2. The FH simulations clearly do not conform to a Poisson distribution, which is manifested in the histograms with a distribution that looks Gaussian at short passage times and with a long decaying tail for longer passage times, and in the cumulative distribution as an overall sigmoidal shape to the kinetic profile. The cumulative

Table 1. Kinetic fits to the cumulative distribution of a sequential process using the functional form of Eq. 6

Protocol	τ_{relax}	σ	τ_1	τ_2	A_1	τ_{single}	$\tau_{\mathrm{stretched}}$	β
FH	350	170	670	9,180	0.5			
FH-visc	317	144	710	6,895	0.4			
FH-fast 10						28,892		
DC-DSE						11,194	15,991	0.84
DC-estimate						7,342	9,595	0.85

$$\Pr(u + j \le t) = \frac{1}{2}\left[1 + erf\left(\frac{t - \tau_{\mathrm{relax}}}{\sigma\sqrt{2}}\right)\right] - \frac{A_1}{2}\left[1 + erf\left(\frac{t - B_1}{\sigma\sqrt{2}}\right)\right]e^{-t/\tau_1}e^{-\frac{C_1}{2\sigma^2}}$$

$$-\frac{(1 - A_1)}{2}\left[1 + erf\left(\frac{t - B_2}{\sigma\sqrt{2}}\right)\right]e^{-t/\tau_2}e^{-\frac{C_2}{2\sigma^2}} \qquad [6]$$

is described in more detail in the supporting information. Shown are kinetic fits to the cumulative distribution of a sequential process for the FH protocol starting from extended state configurations (FH) and for the high-viscosity data (FH-visc). The DC from estimated DSE (DC-estimate) and DC from true DSE (DC-DSE) are fit to both single exponential and stretched exponential functional forms. We also estimate the rate from the FH protocol based on the slope from data of the fastest ≈10 folding trajectories (FH-fast 10) as per Eq. **2**. These should be compared with the equilibrium folding rate τ_f = 15,700.

distribution for the kinetic data are well fit to a sequential process involving a lag phase and diffusion timescale for reaching a collapsed state from the artificially created, completely extended state and a subsequent process involving folding through a set of barriers using a multiple exponential form (see the supporting information and Table 1). We estimate that the first ≈600 and perhaps up to half of the folding trajectories are dominated by the relaxation process and do not experience a significant barrier to folding, a point to which we turn below. To make direct comparisons with previous FH studies, which typically see only on the order of 10 folding events (1), we also fit the first 10 folding events to an assumed single exponential process using only the first-order term in Eq. **2**. The time constant derived from the slope of the first 10 folding events is longer than the equilibrium folding time (Table 1). We also considered artifacts arising from the use of a low-viscosity parameter by running the same FH protocol under the high-viscosity limit; we found no changes in the observed behavior, and it is still well fit to the sequential process of relaxation followed by folding through multiple barrier heights (Table 1).

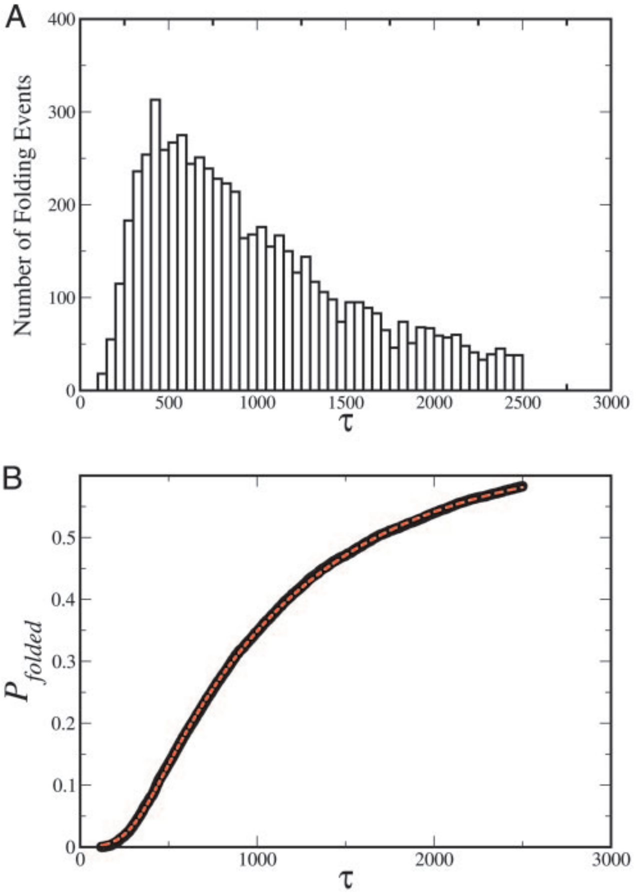

Fig. 2. Distribution of the first passage times for folding (*A*) and plot of percent folded vs. time (*B*) for the FH simulations. (*A*) The non-Poisson behavior is evident in the histogram, with early events appearing to be normally distributed. (*B*) The data (black) are fit well to a sequential process (red) involving relaxation from the fully extended state to a collapsed state, followed by a barrier process involving multiexponential behavior.

To further analyze this kinetic behavior, we look at contact maps for the structures corresponding to the time points just before folding in the FH simulations. These points are taken to be analogous to the folding TSE and can be usefully compared with the TSE evaluated for protein L in previous equilibrium studies using the P_{fold} metric (17). We interpret the non-Poisson distribution for FH trajectories as arising from three different populations. The first population is negligibly small, corresponding to the fastest ≈10–20 first passage times, and shows extremely native-like transitions to folding (Fig. 3*A*). They fold without relaxing to the DSE, do not cross a recognizable barrier, and are unphysical folding trajectories. The remaining population shows evidence of folding through a barrier that resembles the true TSE (Fig. 3*B*), indicating that these trajectories do rapidly fall into the basin corresponding to the DSE. Thus, it is possible for FH simulation to correctly identify folding mechanisms using early folding events, assuming that a good approximation to the true free-energy surface is used.

However, the earliest folders of this second population of folding trajectories represent a biased sampling of Boltzmann states in the DSE (i.e., high-energy states that can more quickly fold depending on their proximity to the barrier and therefore will underestimate the folding rate). This second population can be divided into two subpopulations in which the first corresponds to up to the shoulder region of the FH distribution ($t < 1{,}000\tau$). Contact maps of the fastest folders of this subpopulation at the beginning of their trajectories ($t = 120\tau$, which is somewhere between 25% and 50% of their total folding time) show that they more closely resemble the equilibrated TSE than they do the DSE (Fig. 4*A*) and hence are at the best high-energy states in the DSE. These trajectories represent the features in the cumulative distribution that are dominated by τ_{relax}.

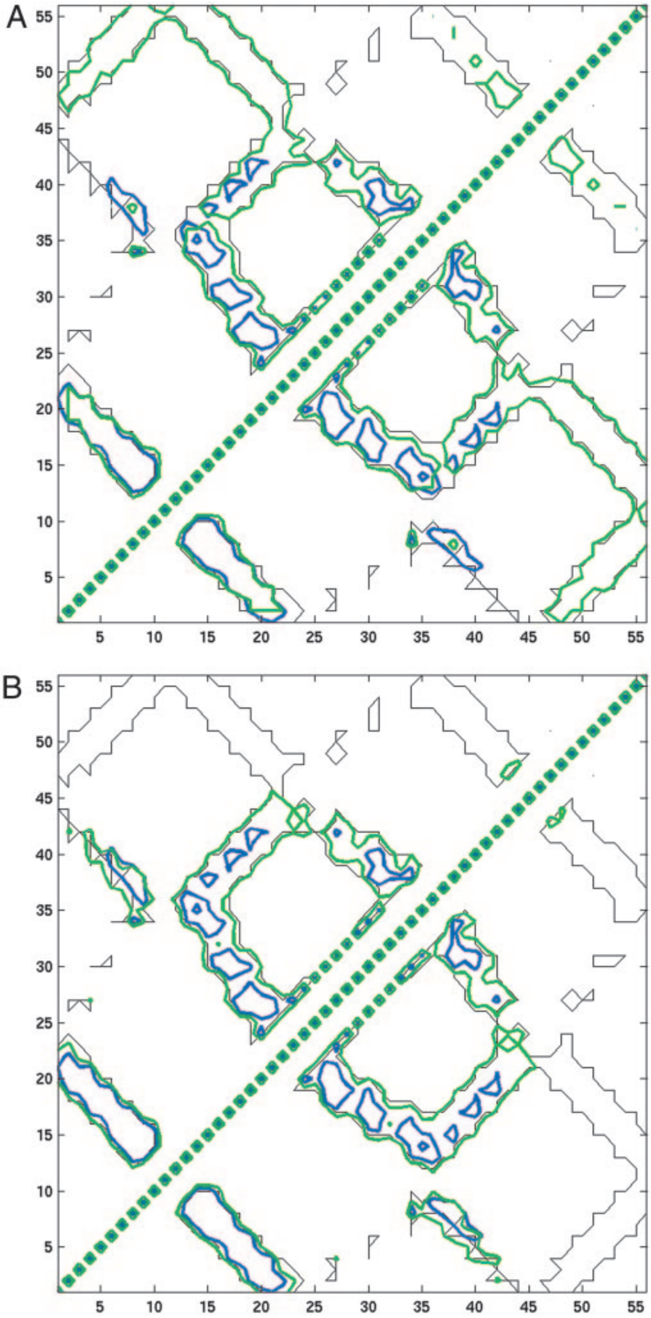

Fig. 3. Contact map of the TSE (blue, contoured at 80% of the population) vs. estimates of the TSE of the 10 fastest folding trajectories (*A*) and the next 100 fastest folding trajectories (*B*) (green, 50%). Native state contacts are contoured in black for comparison. (*A*) The structures are extremely native-like, indicating a barrierless transition to the native state. (*B*) This population experiences a pathway similar to the TSE from the long time equilibrium simulations and is consistent with the remaining FH trajectories.

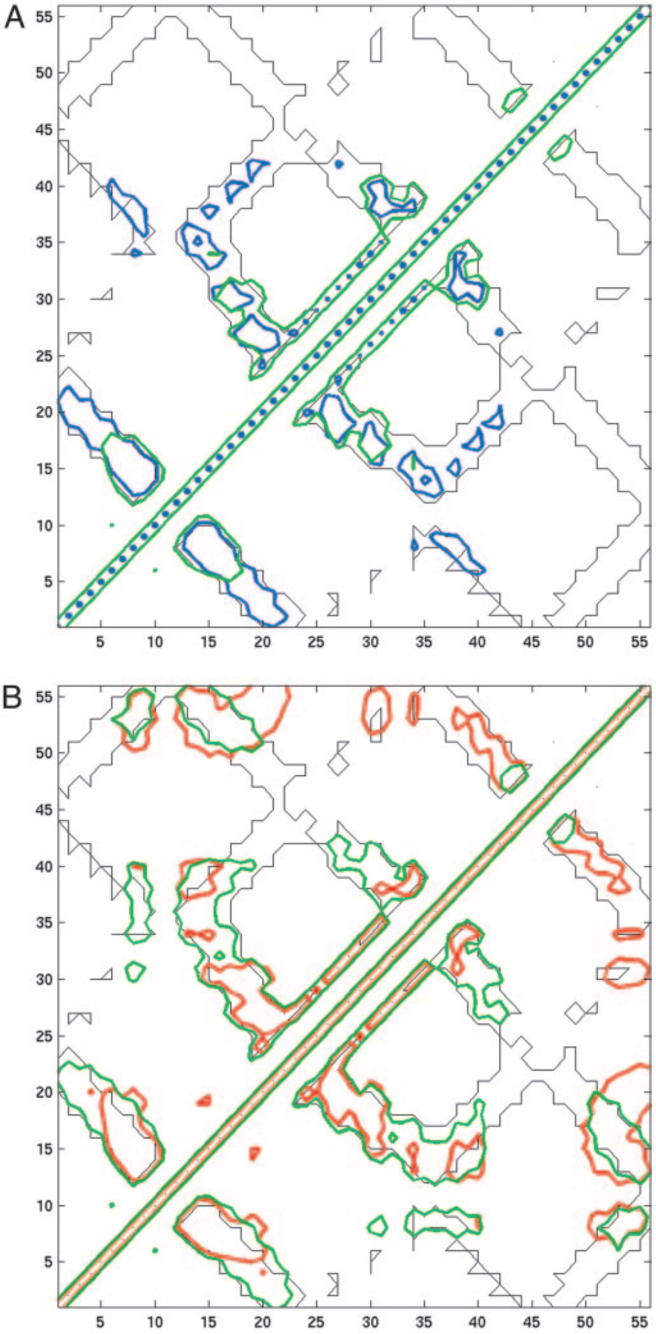

Fig. 4. Contact maps of early time points in folding trajectories. (*A*) Contact map of the TSE (blue, 80%) vs. ensemble of structures of the early time points ($t = 120\tau$) of the faster folding trajectories of the FH simulations (green, 30%). These structures represent the simulations that relax to high-energy states in the DSE and therefore more closely resemble the TSE and not the most probable energy state within the denatured basin. (*B*) Contact maps of the DSE (red, 45%) vs. ensemble of structures of early time points ($t = 500\tau$) of the slower folding trajectories of the FH simulations (green, 30%). These structures are more like the denatured state, indicating that the relaxation time has been exceeded and that a folding barrier is experienced.

This finding is to be contrasted with contact maps of slower folders at the beginning of their trajectories ($t = 500\tau$, which is somewhere between 20% and 50% of their total folding time), which show that they more closely resemble the equilibrated DSE than they do the TSE (Fig. 4*B*) and hence are sampling more probable energy states in the unfolded basin. However, these longer timescale trajectories still do not have sufficient time to partition into a true representation of the equilibrated DSE, because they still underemphasize the nativeness of the second β-hairpin and are missing important regions of stabilizing nonnative interactions. The average potential energy of the equilibrated DSE is 12.5, whereas the average energy of the sampled DSE states of the short trajectories that fold under the FH protocol is 17.1, which shows that this folding population starts from a destabilized, unfolded state that effectively decreases the rate-limiting barrier. Whereas the longer

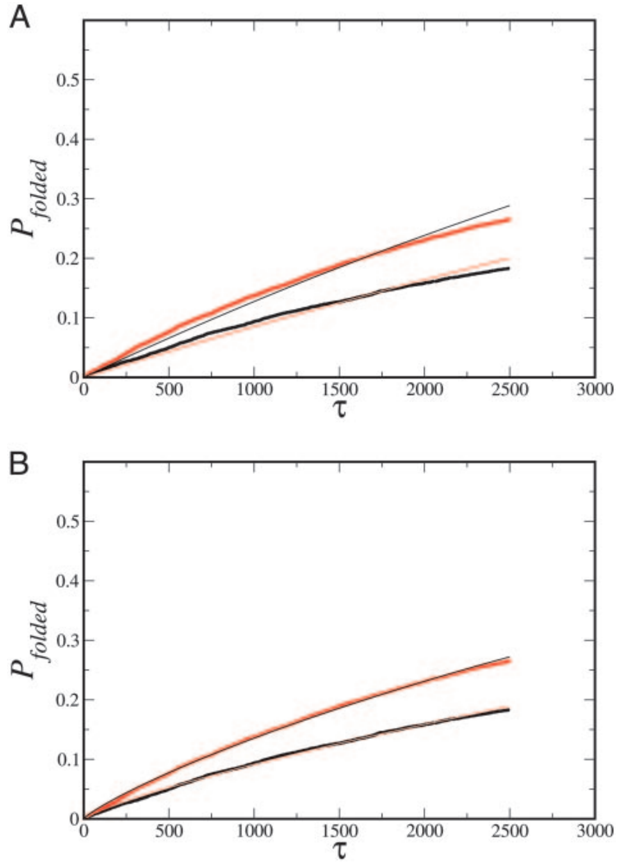

Fig. 5. Percent folded vs. time for simulations launched for the DSE generated from the nonfolding events in the FH trajectories (red) and those from the DSE generated from the long time equilibrium simulations (black). (*A*) The fits to single exponential function are shown, demonstrating that more complex kinetic descriptions are required. (*B*) Same as in *A*, but the data now fit to a stretched exponential, indicating that the simulation time τ_{sim} is too short to represent folding from the deeper energy states in the DSE.

timescale folding population ($>$1,000τ) experiences folding to the native state through a correct barrier type, the corresponding kinetic profile involves the appearance of multiple barriers (depending on which subpopulations in the DSE are sampled in τ_{sim}), and the double exponential fit yields an average rate that is still too fast relative to our equilibrium folding pathway simulation.

In Fig. 5*A*, we show the cumulative distribution for DC uncoupled trajectories simulations launched from the estimated DSE (the FH simulations that did not fold) and the equilibrated DSE, along with kinetics fits to a single exponential process (parameters are summarized in Table 1). The first thing to note is that both simulations underestimate the equilibrium folding time, $\tau_{\mathrm{f}} = 15{,}700$, with the estimated DSE giving a folding time constant of $\tau_{\mathrm{f}} = 7{,}342$ and the equilibrated DSE yielding a time constant of $\tau_{\mathrm{f}} = 11{,}194$. It is also evident that neither simulation fits well to a single exponential; instead, the data are better fit to stretched exponentials as shown in Fig. 5*B* (fit parameters are in Table 1). Although the average folding time constant is $\tau_{\mathrm{f}} = 15{,}991$ from the DSE population, close to the equilibrium folding time, the stretched exponential form indicates that the simulation time τ_{sim} is too short to see the dominance of folding events from the deeper energy states in the DSE.

Fig. 6 shows this bias; we present the distribution of χ values for states of the entire DSE and compare them with the distribution of χ value start states of the 500 fastest folders. As is clear from the histogram, there is a shift to higher χ values for the 500 faster folders, indicating that these starting structures are more native-like, with a skewed sample of strongly native start states (4.6% vs. 0.1% of the DSE). This finding goes along with the assertion that the fastest folding trajectories come from a biased subpopulation in the DSE that has more native-like features resembling the TSE.

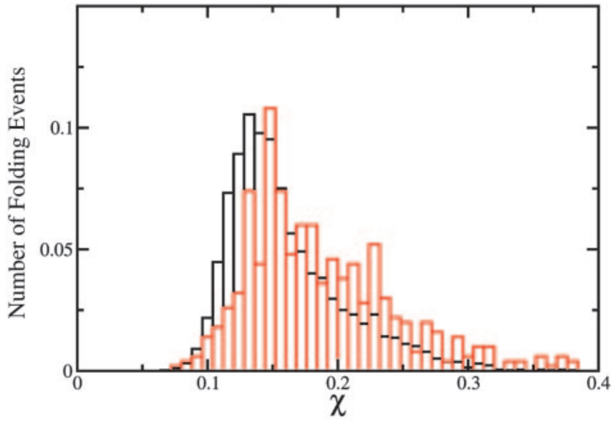

Fig. 6. Histograms showing the distribution of χ values for the entire denatured state (black) and for the denatured state start configurations for the 500 fastest folders (red). A shift to more native-like χ values for the fastest folders is evident, especially in the native region tail of the distribution.

Discussion

The FH approach has been used to predict the kinetic rate, and in most cases some mechanistic observations of what might be a rate-limiting step, for the folding of five small peptides and proteins, including three Trp-zipper peptides (3), a Trp-cage peptide (4), the C-terminal β-hairpin of protein G (1), the designed BBA5 protein (2), and the villin headpiece (HP35) (5). It should be noted that studies on the β-hairpin of protein G used coupled trajectories instead of the uncoupled approach (1). In all cases, the agreement with experimental kinetic rates is excellent, although the simulation error bars reported have significant uncertainty. There is also disagreement with experiment in regards to mechanism of folding for HP35, because purported critical residues found with simulation do not change the experimental rate when mutated (9).

The FH methodology calculates rates from a small number of folding events. In studies on the three Trp-zipper peptides, 150 folding events were observed for TZ1, 212 folding events were observed for TZ2, and 48 folding events were observed for TZ3 (3), and it appears that these events were binned to yield the kinetic data described in the supporting information (3). Eight folding events were used to determine the folding time for the C-terminal β-hairpin from protein G and the Trp-cage peptide (1, 4). In studies on the designed BBA5 peptide, at least 16 folding events were used to determine the folding time (2), and for studies on the villin subdomain, the folding time was estimated from 35 folding events. However, the majority of the 35 folding events of the villin study actually populate a misfolded state, which has the same rms deviation metric as the native state, but nonnative tertiary interactions (5). Zagrovic *et al.* (5) state that if a different structural metric were used, then the folding time they determined would be different. An alternative procedure to help overcome these structural ambiguities has been proposed (18).

Our first observation from our FH simulations is that the first 10–20 trajectories are unphysical, hopping into the native basin without experiencing the rate-limiting barrier. The slope of the cumulative distribution in this region gives a rate constant that is too slow by a factor of two with respect to our experimental benchmark, and because no barrier or alternative barriers to the true TSE are sampled, these earliest of folding events will be unlikely to get the mechanism right. We attribute the apparently slow rate in this region to the distribution of relaxation times from the extended state, which is a broad distribution with a standard deviation of the same order of magnitude as the relaxation time itself. In this initial region, only a fraction of the trajectories have relaxed from the extended state to the DSE, resulting in a slow apparent folding time. Hence, linear kinetic fits in this initial region, which ignore the significant variance of relaxation times, could lead to exaggeratedly slow kinetic rates. If these earliest of folding events are ignored, however, then the next fastest folding events have sufficient time to drop into the DSE to surmount a barrier consistent with the true TSE and therefore are informative about the folding mechanism, although the kinetic profile does not conform to a Poisson process.

The FH protocol using uncoupled trajectories assumes that the relaxation time for the extended state or random-walk state to reach the DSE (t_{relax}), the time to cross the barrier from the DSE to the native state (t_{cross}), the simulation time (t_{sim}), and the experimental folding time (t_{f}), as quoted in ref. 3, will obey the following inequality:

$$t_{\mathrm{relax}} + t_{\mathrm{cross}} < t_{\mathrm{sim}} < t_{\mathrm{f}}. \qquad [4]$$

An important component of the work shown here is that the FH protocol underestimates the relaxation time from the extended state. Although the underlying models and proteins are very different, and although our work has the advantage that our FH simulations exactly match the free-energy surface and native basin definition of the "experiment," we believe that their reported folding rates are very likely in significant error due to $t_{\mathrm{relax}} > t_{\mathrm{sim}}$, at least for the larger proteins such as villin. Although protein L is a longer chain and has more complicated fold topology than the Trp and β-hairpin peptides and small protein BBA5 studied by Pande and coworkers (2, 3), and although it might be argued that relaxation times are comfortably faster than t_{sim} for these smaller peptides and proteins, we are working with a minimalist model of protein L that defines a much smoother free-energy surface than their all-atom model. Furthermore, nearly half of the trajectories are largely experiencing a relaxation process vs. a genuine folding barrier, suggesting that even hundreds of successful folding events in the reported all-atom FH simulations would be insufficient for measuring a true rate. However, the FH study of Trp zippers using both very short and relatively long trajectories found little difference in rate, which suggests that $t_{\mathrm{relax}} < t_{\mathrm{sim}}$ for this system (3), as does recent work on Markovian model formulation using both short and long uncoupled simulations starting from an extended state as input to the state space propagator (7). In cases where folding is studied under high-denaturing conditions, the unfolded basin may be relatively unstructured, and the extended or random-walk state may be a more appropriate estimate of the DSE (19).

A surprising result here is that the DC approach starting from the equilibrated DSE still does not conform to a Poisson distribution (i.e., even if $t_{\mathrm{relax}} < t_{\mathrm{sim}}$, it will still underestimate the rate). In fact, folding trajectories that are too short will always have an inherent bias due to sampling of folding events from states in the DSE that sit closer to the barrier. That is, an additional requirement beyond Eq. **4** is that the time to sample folding events from the deeper regions of the DSE basin, t_{DSE}, be shorter than t_{sim}: i.e.,

$$t_{\mathrm{relax}} + t_{\mathrm{cross}} + t_{\mathrm{DSE}} < t_{\mathrm{sim}} < t_{\mathrm{f}}. \qquad [5]$$

Sampling folding events originating from the most probable regions of the DSE basin may be sufficiently fast for simple peptides and proteins, so that the simulation time can be realistically lengthened on a DC environment so that $t_{\mathrm{DSE}} < t_{\mathrm{sim}}$, and, hence, the complexity of the DSE will not significantly alter the single exponential character of the folding time distribution (19, 20). Thus, the DC paradigm using uncoupled trajectories can match long-timescale ensemble experiments when trajectories are launched from an equilibrated DSE, the mechanism of folding involves a single barrier, and we fold from states of the DSE that are most probable. To make contact with experiment, an additional critical requirement is that the simulated free-energy surface be correct.

Computing architectures based on loosely coupled processors such as FH have exploited trivial parallelization advantages of adequate sampling of the handful of rare and rapid folding events based on short trajectories. However, for more complex proteins that fold by means of a two-state mechanism, ultrashort uncoupled trajectories may significantly alter the single exponential character of the folding time distribution or give a faster timescale for folding, which is exactly what we found from our DC results. In this case, the DC software must be optimized to significantly increase the timescale of "short" trajectories to observe folding from states of the DSE that are most probable to be made more accurate for more complex models or larger protein systems. A future software engineering challenge may be to adapt such DC platforms to use nontrivial parallelization strategies for energy and force interactions where communication bottlenecks must be overcome without a real high-speed interconnect between processors.

Although it is clear that significant improvements in the FH simulation protocol are required and that perhaps the DC approach will ultimately require much longer trajectories than originally hoped for, it has been a very worthwhile effort in that it has revealed our incomplete investigation of the simple two-state approximation of the folding of small proteins. In our view, a challenging direction in protein-folding research is characterization of structure in the denatured state basin because poor sampling statistics of its subpopulations can significantly alter the kinetic profile. In fact, fast time-resolved experiments and single-molecule folding studies may also be picking up on these subpopulations in the DSE (21, 22). This behavior has also been observed in various experimental and simulation studies on helical peptides (23–25). The DC paradigm using short uncoupled trajectories, together with fast time-resolved experiments and single-molecule studies, can better reveal the breakdown in the two-state approximation due to a biased observation of folding from higher energy subpopulations in the DSE.

We thank Vijay Pande and Bill Swope for useful discussions and Dean Scrimgeour for useful discussions on sequential processes. This work was supported by University of California, Berkeley; a subcontract award under National Science Foundation Grant CHE-0205170; and a graduate research fellowship from the Whitaker Foundation (to N.L.F.). T.H.-G. is a Schlumberger Fellow.

1. Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J. & Zagrovic, B. (2003) *Biopolymers* **68,** 91–109.
2. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002) *Nature* **420,** 102–106.
3. Snow, C. D., Qiu, L., Du, D., Gai, F., Hagen, S. J. & Pande, V. S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 4077–4082.
4. Snow, C. D., Zagrovic, B. & Pande, V. S. (2002) *J. Am. Chem. Soc.* **124,** 14548–14549.
5. Zagrovic, B., Snow, C. D., Shirts, M. R. & Pande, V. S. (2002) *J. Mol. Biol.* **323,** 927–937.
6. Voter, A. F. (1998) *Phys. Rev. B Condens. Matter* **57,** R13985–R13988.
7. Singhal, N., Snow, C. D. & Pande, V. S. (2004) *J. Chem. Phys.* **121,** 415–425.
8. Fersht, A. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14122–14125.
9. Kubelka, J., Eaton, W. A. & Hofrichter, J. (2003) *J. Mol. Biol.* **329,** 625–630.
10. Paci, E., Cavalli, A., Vendruscolo, M. & Caflisch, A. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8217–8222.
11. Swope, W. C., Pitera, J. W. & Suits, F. (2004) *J. Phys. Chem. B* **108,** 6571–6581.
12. Brown, S., Fawzi, N. J. & Head-Gordon, T. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10712–10717.
13. Brown, S. & Head-Gordon, T. (2004) *Protein Sci.* **13,** 958–970.
14. Fawzi, N. L., Chubukov, V., Clark, L. A., Brown, S. & Head-Gordon, T. (2005) *Protein Sci.* **14,** 993–1003.
15. Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R., III, & Baker, D. (1997) *Biochemistry* **36,** 3373–3382.
16. Andersen, H. C. (1983) *J. Comput. Phys.* **52,** 24–34.
17. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998) *J. Chem. Phys.* **108,** 334–350.
18. Marianayagam, N. J., Brown, A. G. & Jackson, S. E. (2005) *J. Biomol. Struct. Dyn.* **23,** 73–76.
19. Klimov, D. K., Newfield, D. & Thirumalai, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8019–8024.
20. Krivov, S. V. & Karplus, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 14766–14770.
21. Yang, W. Y. & Gruebele, M. (2003) *Nature* **423,** 193–197.
22. Ma, H. & Gruebele, M. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 2283–2287.
23. Hummer, G., Garcia, A. E. & Garde, S. (2000) *Phys. Rev. Lett.* **85,** 2637–2640.
24. Huang, C. Y., Klemke, J. W., Getahun, Z., DeGrado, W. F. & Gai, F. (2001) *J. Am. Chem. Soc.* **123,** 9235–9238.
25. Sorin, E. J. & Pande, V. S. (2005) *Biophys. J.* **88,** 2472–2493.